Series 5

1. **Model diagnostics: simulation study** Assessing model diagnostic plots requires experience. Often it is difficult to decide whether a deviation from the theoretical centre is a systematic one (i.e. needing correction) or a random one (i.e. just variability in the data). Experience can be gained by performing model diagnostics on problems where it is known whether the model assumptions hold or do not hold. This allows us to identify the naturally occurring variability in the results.

   Simulate the following 4 models: one of them fulfils all model assumptions, other includes a systematic deviation from the linearity assumption,e.g., $\mathbb{E}[\epsilon_i] \neq 0$, and the two left include minor and major deviations from the constant variance assumption.

```
> n <- 100
> xx <- 1:n
> yy.a <- 2+1*xx+rnorm(n)
> yy.b <- 2+1*xx+rnorm(n)*(xx)
> yy.c <- 2+1*xx+rnorm(n)*(1+xx/n)
> yy.d <- cos(xx*pi/(n/2)) + rnorm(n)
```

   a) Decide which model has no violation of the model assumptions, minor deviation to non-constant variance, major deviation to non-constant variance and which model is non-linear.

   b) Plot each response `yy.[a,b,c,d]` versus `xx`. Fit a simple linear regression and plot the regression line into the corresponding scatter plot.

   c) Perform model diagnostics and have a look at the diagnostic plots. Where can we see the deviations? How large is the random variation within these plots?

   d) Repeat generating the random numbers a few times and study the variation in the resulting plots. You can also change the number of observations and track the changes in the plots.

   Assessing normal plots is equally difficult[1]. Even drawing samples from a normal distribution does not result in observations lying directly on the straight line. Now, we will use to following code to simulate new data and see how a skewed, a long-tailed and a short-tailed distribution look like in a Q-Q plot:

```
> qqnorm(rnorm(n), main=c("Normal distribution"))
> qqnorm(exp(rnorm(n)), main=c("Lognormal distribution"))
> qqnorm(rcauchy(n), main=c("Cauchy distribution"))
> qqnorm(runif(n), main=c("Uniform distribution"))
```

   e) Decide which random numbers are normal, skewed, short-tailed and long-tailed.

   f) Repeat generating the random numbers a few more times and study the variation in the resulting Q-Q plots. You can also change the number of observations and track the changes in the plots.

2. **Regression with a factor**
   Sometimes, there are continuous and categorical independent variables. The categorical variable is sometimes called factor. We consider two predictors: a continuous variable and a categorical variable that has K factors/levels. We denote the continuous variable by z and introduce the following variables:

$$x^{(j)} = \begin{cases} 1 & \text{if category} = j \\ 0 & \text{else} \end{cases}$$

[1] In the StatsNotes of the Department of Mathematics and Statistics at Murdoch University it reads: *A sufficiently trained statistician can read the vagaries of a Q-Q plot like a shaman can read a chicken's entrails, with a similar recourse to scientific principles. Interpreting Q-Q plots is more a visceral than an intellectual exercise. The uninitiated are often mystified by the process. Experience is the key here.*

for $j = 2, \ldots, K$. The linear model for the observations $i = 1, \ldots, n$ is then

$$y_i = \alpha + \beta z_i + \sum_{j=2}^{K} \gamma_j x_i^{(j)} + \sum_{j=2}^{K} \delta_j z_i x_i^{(j)} + \varepsilon_i.$$

The errors ε_i are assumed to be i.i.d. $\sim \mathcal{N}\left(0, \sigma^2\right)$.

a) Give an interpretation of the model and the coefficients. What are the differences to a simple linear model for every single category?

b) Give a mathematical formulation for the null hypothesis "the effect of an increase of z by one unit is the same for all categories".

(i) Write down the test statistic that can be used to test the above hypothesis.

(ii) What is the distribution of the test statistic if the null hypothesis is true?

c) Perform a test of the above hypothesis for an example: The data set `salary.dat` contains a dependent variable `y` describing the annual income of an employee. The continuous independent variable is `experience` (work experience in years). The factor is `education` (the categories are "Berufslehre", "Maturität" ad "Hochschulabschluss" describing different levels of education).

d) Verify that the bigger model satisfies the model assumption better than the smaller model by comparing their Tukey-Anscombe plots and QQ-plots.

R-hints: Data import:

```
salary <- read.table("http://stat.ethz.ch/Teaching/Datasets/salary.dat", header=TRUE)
```

R interprets the values in `salary$education` as integers. We have to make it a factor with the command

```
salary$education <- as.factor(salary$education)
```

and the general linear model is then fitted with

```
fit1 <- lm(y ~ experience + education + experience:education, data=salary).
```

The reduced model, where the effect of an increase of `experience` is the same for all levels of education, is fitted with

```
fit2 <- lm(y ~ experience + education, data=salary).
```

The p value of the test is obtained with

```
anova(fit2, fit1).
```

3. Consider a multiple linear regression model, where the parameters are estimated using the least square method. Choose the correct answer(s).

a) The sum of the residuals can be non-zero

(i) if the model contains at least one categorical variable.

(ii) if the "zero-mean" assumption is not correct.

(iii) if no predictor is significant.

(iv) if the model does not contain an intercept term.

(v) under no circumstances.

b) The global F-test produces a very small p-value but none of the predictors are significant. This situation can occur

(i) if the predictors are correlated.

(ii) if the response variable is independent of all the predictors.

(iii) under no circumstances.

Preliminary discussion: Monday, November 10.

Deadline: Monday, November 17.